Available online at www.sciencedirect.com
SCIENCE DIRECT
Remote Sensing of Environment 86 (2003) 500–516
Remote Sensing of Environment
www.elsevier.com/locate/rse

Thematic accuracy of the 1992 National Land-Cover Data for the eastern United States: Statistical methodology and regional results

S.V. Stehman[a,*], J.D Wickham[b], J.H. Smith[c], L. Yang[d]

[a] *SUNY, College of Environmental Science and Forestry, 320 Bray Hall, Syracuse, NY 13210, USA*
[b] *Landscape Characterization Branch, U.S. EPA, E243-05, Research Triangle Park, NC 27711, USA*
[c] *National Center for Environmental Research, U.S. EPA, 1200 Pennsylvania Ave. NW, Washington, DC 20460, USA*
[d] *Science Applications International Corporation (SAIC), EROS Data Center, Sioux Falls, SD 57198, USA*

Received 9 August 2002; received in revised form 24 April 2003; accepted 3 May 2003

Abstract

The accuracy of the 1992 National Land-Cover Data (NLCD) map is assessed via a probability sampling design incorporating three levels of stratification and two stages of selection. Agreement between the map and reference land-cover labels is defined as a match between the primary or alternate reference label determined for a sample pixel and a mode class of the mapped 3 × 3 block of pixels centered on the sample pixel. Results are reported for each of the four regions comprising the eastern United States for both Anderson Level I and II classifications. Overall accuracies for Levels I and II are 80% and 46% for New England, 82% and 62% for New York/New Jersey (NY/NJ), 70% and 43% for the Mid-Atlantic, and 83% and 66% for the Southeast.
© 2003 Elsevier Inc. All rights reserved.

Keywords: National Land-Cover Data; Thematic accuracy; Eastern United States

1. Introduction

The Multi-Resolution Land Characteristics Consortium (MRLC) sponsored the production of the 1992 National Land-Cover Data (NLCD) of the conterminous United States (Loveland & Shaw, 1996). The NLCD was created from early 1990s Landsat Thematic Mapper (TM) imagery, augmented by a suite of geospatial ancillary data. It displays 21 thematic classes employing a modified Anderson, Hardy, Roach, and Witmer (1976) land use and land cover classification system (Vogelmann et al., 2001). The 1992 NLCD represents the first continuous coverage, seamless land-cover map produced for the conterminous United States. The accompanying accuracy assessment represents one of the first efforts to evaluate the accuracy of a national land-cover map employing a detailed classification scheme (Anderson Level II) at a 30-m pixel resolution.

NLCD construction is organized by 10 EPA Federal administrative regions. The accuracy assessment is similarly organized, with implementation and analysis conducted separately in each region. The four regions comprising the eastern U.S. are New England, New York/New Jersey (NY/NJ), Mid-Atlantic, and Southeast (Fig. 1, Table 1). Although some regional variation in protocol and implementation exists, the accuracy assessment strategies employed in each region share a common general framework. The sampling design criteria applied to all four regions are that the design should: (1) satisfy protocols defining a probability sample; (2) provide adequate sample sizes for estimating user's accuracies with acceptable precision; (3) be cost efficient regarding materials (e.g., aerial photography) required to obtain the reference classification; and (4) achieve a spatially well-distributed sample.

The objectives of this article are to document the accuracy assessment protocol and to report accuracy results for the four regions of the eastern U.S. Yang, Stehman, Smith, and Wickham (2001) provide a condensed overview of this methodology and report accuracy aggregated over all four regions. Other portions of the NLCD accuracy assessment

* Corresponding author. Tel.: +1-315-470-6692; fax: +1-315-470-6535.
E-mail address: svstehma@syr.edu (S.V. Stehman).

0034-4257/03/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0034-4257(03)00128-7

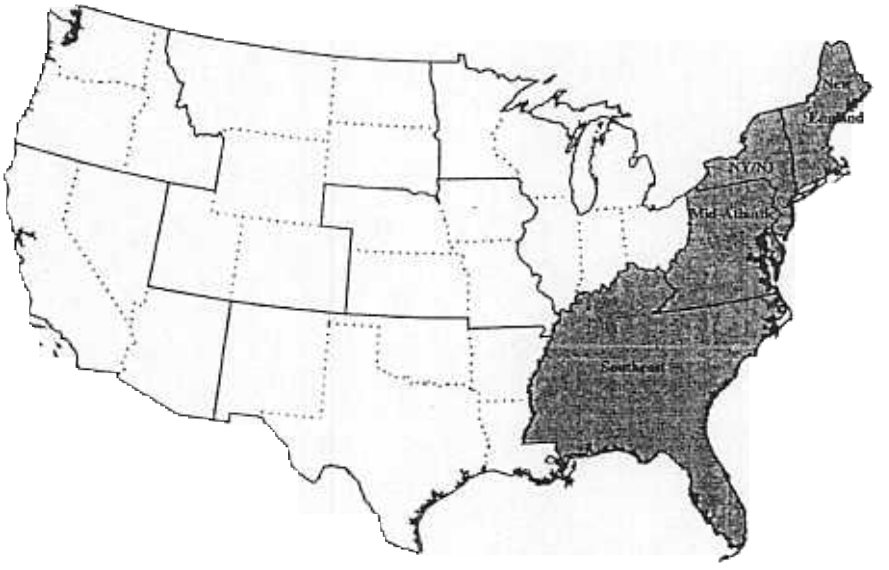

Fig. 1. Location map. Federal region boundaries are depicted using solid lines and state boundaries are depicted using dotted lines. The eastern U.S. study area is shown in gray.

methodology have been described by Stehman, Wickham, Yang, and Smith (2000), Zhu, Yang, Stehman, and Czaplewski (1999, 2000), and Yang, Stehman, Wickham, Smith, and VanDriel (2000).

Table 1
Regional distribution of mapped land-cover (% of area) for the 1992 NLCD Level II classification

Class	New England	New York/ New Jersey	Mid-Atlantic	Southeast
11 Water			6.485	6.882
21 Low-density residential			2.520	1.968
22 High-density residential	0.378	1.232	0.339	0.618
23 Commercial/ industrial	1.249	1.083	0.521	0.792
31 Bare rock/ sand/clay	0.198	0.048	0.014	0.073
32 Quarries/ strip mining	0.064	0.129	0.470	0.104
33 Transitional land	1.238	0.070	0.521	2.284
41 Deciduous forest	27.122	31.882	47.292	22.961
42 Evergreen forest	17.112	5.464	5.529	17.206
43 Mixed forest	24.053	16.405	9.968	11.428
51 Shrubland	0.264	0.000	0.000	0.000
81 Pasture/hay	1.742	9.198	18.623	10.337
82 Row crops	4.741	12.237	5.067	13.331
85 Urban/ recreational grasses	0.911	0.748	0.142	0.419
91 Woody wetlands	3.401	2.906	.683	9.552
92 Emergent wetlands	.268	0.716	0.826	2.045

2. Methods

The description of the NLCD accuracy assessment methodology focuses on three primary components (Stehman & Czaplewski, 1998): (1) the sampling design, which determines the spatial locations at which the reference data are obtained; (2) the response design, which details how the reference data are obtained; and (3) the analysis plan for producing the accuracy estimates. Each component will be described in detail in subsequent sections. Because the 1992 NLCD final product is delivered unfiltered and unsmoothed at a 30-m pixel resolution, the accuracy assessment methodology is applied to this 30-m product.

Describing the accuracy of the 1992 NLCD map is the primary objective of this assessment. Description focuses on the error matrix and accompanying summary measures including overall, user's and producer's accuracies. Standard errors for these estimated summary measures are also reported. Error matrices are constructed with the rows representing the map land cover and the columns representing the reference land cover. Accuracy results are reported for each of the four eastern mapping regions at both Levels I and II.

3. Sampling design

The sampling design employed a nested, hierarchical partition of the eastern U.S. constructed of spatial units of four different sizes: 30 m pixels, primary sampling units (PSUs) which were clusters of pixels, geographic strata containing the PSUs, and mapping regions encompassing the geographic strata (Fig. 2). Each of the four mapping regions constituted a separate sampling objective, and these

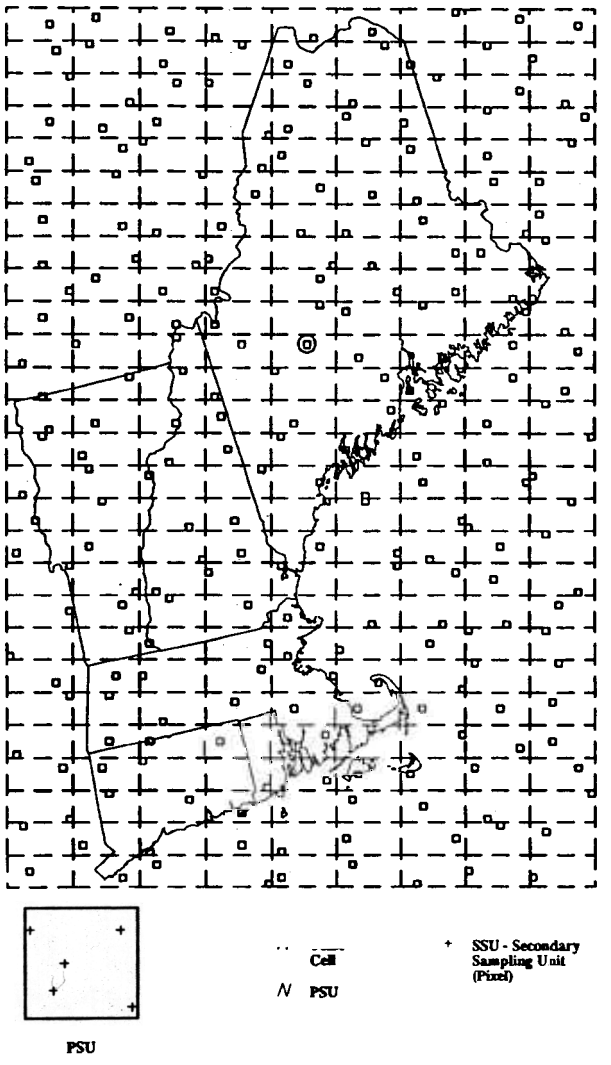

Fig. 2. Sampling structure for New England region. The geographic strata are shown using a dashed line. Selected PSUs are shown using a thick solid line and state boundaries are shown using a thin solid line. The circled PSU in south–central Maine is "exploded" to show the SSUs, the pixels.

regions represented the first level of stratification of the design. The next level of stratification employed geographic strata consisting of the PSUs within each of the four regions. In three of the mapping regions, these geographic strata were 15′ by 15′ sections and, in the larger Southeast region, 30′ by 30′ sections were employed. Each geographic stratum was further subdivided into an equal number of PSUs, each PSU approximately 6 by 6 km. The PSUs formed a tessellation obtained by a tiling of NAPP photographs in which only interior, non-overlapping areas of the photographs were used. Each PSU was a cluster of pixels, and a pixel was the ultimate or secondary sampling unit.

In the first stage of the sampling protocol, one PSU was randomly selected from each geographic stratum. Because all geographic strata have the same number of PSUs, each PSU in a mapping region had the same probability of being sampled. For geographic strata extending outside the mapping region, PSUs not in the mapping region remained eligible to be selected so that the number of PSUs was equal for all geographic strata. However, if a PSU outside of the mapping region was selected, no sample data were collected for that PSU. Pixels within the sampled first-stage PSUs were next stratified by mapped Level II land-cover class. At the second-stage, a simple random sample of pixels was then selected from the first-stage sample pixels within each land-cover class stratum. The target sample size for each stratum was 100 pixels.

The NY/NJ region was the first region assessed. Data were collected from two sampling designs (Zhu et al., 2000). For the "general" design, PSUs were organized into geographic strata and the first-stage selection protocol common to all mapping regions was implemented. However, stratification by land-cover class was not employed at the second stage in the NY/NJ general design. Instead, sample pixels were selected without regard to land-cover class, with 4 pixels targeted per PSU. Sample sizes for the rare map classes were expected to be small, and user's accuracy estimated imprecisely with only these data. Consequently, a second design was implemented to augment the following rare classes: high-density residential (22), commercial/industrial (23), bare rock/sand/clay (31), quarries/strip mine (32), transitional land (33), urban/recreational grasses (85), and emergent wetland (92). This "rare class" design followed the two-stage, stratified protocol subsequently employed in the other three mapping regions. A difference between the results reported here for the NY/NJ region and those presented in Zhu et al. (2000) is that the data from the two samples are now combined in the accuracy estimates. In the earlier work, results were reported separately for the different sampling designs. In all subsequent regions assessed, the initial "two-design" approach implemented in NY/NJ was modified in favor of a simpler, single design in which stratification was employed at the second stage of the selection protocol.

In summary, the sampling design can be characterized as having three levels of stratification, mapping region, tessellation cell (i.e., collection of PSUs), and map land-cover class, and two stages of selection, the first-stage selection of PSUs followed by a second-stage selection of pixels. The sampling protocol satisfied the four desirable design criteria specified in Section 1. Stratifying by map land-cover class was employed to ensure adequate precision for estimating user's accuracies. Restricting the sample pixels to the first-stage sample PSUs reduced the number of air photos that had to be purchased, thus achieving the desired reduction in reference material cost. Without this spatial clustering, sample pixels would have been more geographically dispersed requiring many more photos to obtain the reference data. The geographic stratification improved the spatial distribution of the sample relative to the distribution of PSUs selected via simple random sampling. Lastly, all phases of the design were constructed to satisfy criteria defining a probability sample.

The reference land-cover classifications were obtained by photointerpreting 1:40,000-scale National Aerial Photography Program (NAPP) black-and-white or color infrared film acquired during the period 1989–1993, a time span approximately concurrent with the dates of the TM imagery used to produce the 1992 NLCD. Each sample pixel was located on the Landsat three-band composite TM spectral image using the spatial coordinates specified by the sampling design. Locating the sample point on the spectral image instead of the classified image maintained interpreter ignorance of the map land-cover label of the sample pixel. Accurately locating the sample point on the non-georeferenced NAPP prints was aided by visually consulting spatial patterns evident on the TM color composite image (Zhu et al., 2000). Interpreters determined the most likely (i.e., primary) land-cover class for each pixel, and had the option of providing one alternate reference land-cover label. A photointerpreter confidence rating of the assigned primary reference label was included in the response design protocol (Yang et al., 2000). Photointerpretation was conducted by different teams of interpreters for the different mapping regions resulting in some regional variation in the response design. Interpreter training and quality control procedures also varied among regions.

4.1. Defining agreement

The response design protocol requires specifying a definition of agreement when comparing the map and reference classifications. Two imperfections in reference data collection influence this definition: inability to precisely co-register the reference and map locations (i.e., positional uncertainty), and difficulty in assigning a single, crisp reference label to a pixel (i.e., land-cover class ambiguity). Positional and thematic uncertainty are not independent (Lanter & Veregin, 1992), and different definitions of agreement assign different emphasis to these sources of uncertainty in the reference data. For example, defining agreement as a match between the map label and the primary reference label imposes a crisp set reference labeling protocol and accommodates neither positional uncertainty nor thematic ambiguity. Both sources of uncertainty can be factored into the definition of agreement by introducing additional information related to the map attribute, the reference attribute, or both. For example, if both a primary and an alternate reference land-cover label are provided, agreement may be declared if the map class matches either the primary or alternate reference label. This definition accommodates thematic ambiguity of the reference data.

Agreement may also be defined on the basis of a spatial support region (e.g., a 3×3 block of pixels centered on the sample pixel) and either map or reference data contained within this support region. For the NLCD mode definition of agreement, we used the map labels to determine one or more modal land-cover classes for a 3×3 support region and defined agreement as a match between a mode map class and either the primary or alternate reference label of the sample pixel. For this definition, it is possible for several classes to qualify as modes. In such cases, agreement was declared if any map mode class matched either the primary or alternate reference label. If more than one map class qualified as a mode but none matched the primary or alternate reference label, the error matrix row location of that sample pixel was determined by the map label of the center pixel. The definition of agreement based on a 3×3 pixel support region was applied to both Level I and II of the NLCD. When applied to Level I, the pixels in the support region were converted to Level I classes prior to determining the mode map class(es).

5. Analysis

5.1. Estimating accuracy

The analysis is derived from the general estimation theory of probability sampling (cf. Särndal, Swensson, & Wretman, 1992). The analysis requires determining the inclusion probabilities resulting from the sampling protocol (Stehman, 2001; Stehman & Czaplewski, 1998), where an inclusion probability is the probability that a particular pixel is included in the sample. Inclusion probabilities are necessary to construct statistically consistent estimates of accuracy.

The two-stage structure of the sampling design generates inclusion probabilities for each stage. The first-stage inclusion probability, denoted π_{1u}, is determined by the protocol used to select the sample of PSUs. By construction, all geographic strata within a mapping region had the same number of PSUs, K. Consequently, the probability of selecting any particular PSU was $1/K$. Each pixel within a PSU was sampled with the same probability as its containing PSU, so $\pi_{1u} = 1/K$ for each pixel in the mapping region.

At the second stage, those pixels selected in the first-stage sample were stratified by their mapped land-cover class. Consider a particular land-cover class denoted by the subscript h. Suppose N_h^* pixels mapped as class h were contained in the first-stage sample of PSUs. A simple random sample of size n_h of these class h pixels was selected from the N_h^* pixels available. Conditional on the selected first-stage sample, the second-stage inclusion probability for each pixel of class h was $\pi_{2.1hu} = n_h/N_h^*$. Consequently, the

Table 2
Overall accuracy (% ± standard error) by region for Level I and Level II NLCD using the mode definition of agreement

Region	Level II	Level I
New England	46 ± 2.2	80 ± 1.7
New York/New Jersey	62 ± 1.4	82 ± 1.2
Mid-Atlantic	43 ± 3.9	70 ± 2.6
Southeast	66 ± 2.0	83 ± 1.4

Table 3
Error matrix for Level I and Level II classification: New England region

Level II

	21	22	23	31	32	33	41	42	43	51	81	82	85	91	92	Total	Users	S.E.	*n*
9.081	0.000	0.000	0.014	0.019	0.000	0.000	0.003	0.179	0.000	0.000	0.092	0.000	0.000	0.014	0.000	9.402	0.97		113
0.036	1.641	0.154	0.144	0.000	0.000	0.000	0.271	0.744	0.036	0.117	0.116	0.000	0.091	0.000	0.070	3.419	0.48		107
0.007	0.246	0.305	0.134	0.015	0.000	0.000	0.007	0.007	0.000	0.000	0.000	0.022	0.031	0.000	0.000	0.775	0.39		99
0.175	0.372	0.000	0.749	0.046	0.000	0.015	0.051	0.092	0.031	0.015	0.061	0.031	0.123	0.002	0.000	1.764	0.43		105
0.004	0.019	0.000	0.013	0.043	0.000	0.002	0.008	0.004	0.000	0.019	0.008	0.004	0.008	0.000	0.017	0.149	0.29		67
0.002	0.000	0.000	0.007	0.004	0.017	0.002	0.001	0.004	0.001	0.001	0.001	0.006	0.000	0.000	0.001	0.048	0.35		77
0.000	0.014	0.000	0.012	0.066	0.000	0.526	0.037	0.078	0.028	0.169	0.021	0.012	0.012	0.012	0.024	1.014	0.52		97
0.016	0.408	0.000	0.014	0.272	0.009	0.312	12.280	2.947	4.863	1.941	1.775	0.340	0.292	0.072	0.000	25.541	0.48		159
0.357	0.000	0.000	0.473	0.746	0.000	0.685	1.951	7.589	4.851	1.495	0.473	0.012	0.009	0.629	0.470	19.740	0.38		137
0.068	0.036	0.000	0.061	0.566	0.000	0.270	4.573	8.647	10.151	1.006	0.513	0.000	0.009	0.084	0.017	26.002	0.39		135
0.000	0.003	0.000	0.000	0.007	0.000	0.027	0.000	0.007	0.003	0.162	0.003	0.000	0.000	0.000	0.003	0.217	0.75		58
0.000	0.206	0.000	0.000	0.039	0.000	0.019	0.039	0.378	0.039	0.039	0.977	0.246	0.019	0.000	0.019	2.021	0.48		90
0.067	0.172	0.000	0.133	0.082	0.104	0.314	0.208	0.595	0.158	0.212	1.565	1.796	0.086	0.052	0.000	5.544	0.32		114
0.018	0.086	0.000	0.046	0.019	0.000	0.009	0.064	0.065	0.009	0.003	0.104	0.091	0.161	0.000	0.009	0.686	0.24		68
0.036	0.014	0.034	0.063	0.016	0.014	0.034	0.308	0.411	0.406	0.151	0.001	0.000	0.000	0.723	0.498	2.708	0.27		88
0.055	0.052	0.014	0.000	0.000	0.000	0.014	0.233	0.116	0.000	0.068	0.000	0.023	0.027	0.055	0.315	0.971	0.32		59
9.924	3.270	0.506	1.864	1.940	0.144	2.231	20.033	21.863	20.576	5.399	5.710	2.584	0.869	1.643	1.445	100.001			
0.92	0.50	0.60	0.40	0.02	0.12	0.24	0.61	0.35	0.49	0.03	0.17	0.70	0.19	0.44	0.22				
0.03	0.06	0.11	0.08	0.01	0.09	0.06	0.05	0.04	0.05	0.01	0.03	0.10	0.07	0.10	0.10				
134	158	44	108	71	31	91	136	198	113	132	124	74	49	46	64				1573

Level I

	20	30	40	50	80	90	Total	Users	S.E.	
9.081	0.014	0.019	0.182	0.000	0.092	0.014	9.402		0.021	113
0.218	3.724	0.086	0.911	0.061	0.392	0.036	5.427		0.034	300
0.006	0.066	0.567	0.098	0.172	0.054	0.053	1.016		0.057	212
0.456	0.966	2.993	59.711	4.232	3.713	0.901	72.971		0.021	495
0.000	0.003	0.031	0.010	0.148	0.003	0.003	0.199		0.072	53
0.085	0.703	0.493	1.568	0.433	4.880	0.081	8.243		0.049	274
0.077	0.139	0.075	0.737	0.178	0.030	1.505	2.742		0.058	126
9.924	5.615	4.264	63.218	5.222	9.163	2.593	100.000			
0.92	0.66	0.13	0.95	0.03	0.53	0.58				
0.03	0.06	0.03	0.01	0.01	0.06	0.10				
134	308	188	460	132	247	104				1573

Table 4
Error matrix for Level I and Level II classification: New York/New Jersey region

Level II

Class	11	21	22	23	31	32	33	41	42	43	81	82	85	91	92	Total	Users	S.E.	*n*
	8.374	0.000	0.000	0.000	0.103	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.103	8.581	0.98		85
	0.000	2.708	0.447	0.169	0.000	0.000	0.084	0.022	0.084	0.000	0.000	0.000	0.268	0.000	0.000	3.782	0.72		49
	0.000	0.230	1.208	0.471	0.000	0.000	0.000	0.020	0.136	0.000	0.000	0.000	0.000	0.020	0.020	2.106	0.57		67
	0.022	0.149	0.000	1.463	0.022	0.022	0.043	0.043	0.000	0.000	0.000	0.116	0.043	0.000	0.112	2.035	0.72		58
	0.000	0.084	0.003	0.002	0.054	0.000	0.002	0.000	0.106	0.000	0.000	0.000	0.092	0.002	0.005	0.350	0.15		43
	0.003	0.097	0.003	0.106	0.000	0.257	0.004	0.003	0.003	0.000	0.003	0.003	0.000	0.003	0.000	0.484	0.53		44
	0.000	0.000	0.000	0.005	0.000	0.000	0.026	0.102	0.115	0.100	0.000	0.000	0.000	0.016	0.004	0.369	0.07		42
	0.281	0.190	0.095	0.666	0.000	0.095	0.286	21.606	3.776	2.454	1.899	2.927	0.381	0.593	0.286	35.536	0.61		377
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.096	4.686	0.096	0.000	0.096	0.000	0.096	0.000	5.070	0.92		53
	0.096	0.457	0.000	0.022	0.000	0.000	0.090	1.718	2.734	7.584	0.368	0.456	0.000	0.181	0.181	13.887	0.55		154
	0.000	0.368	0.182	0.368	0.000	0.000	0.000	0.460	0.457	0.184	4.050	3.314	0.184	0.000	0.000	9.568	0.42		104
	0.000	0.279	0.000	0.560	0.095	0.000	0.184	1.124	0.374	0.184	1.761	7.660	0.092	0.000	0.095	12.407	0.62		133
	0.000	0.198	0.015	0.073	0.000	0.000	0.000	0.099	0.000	0.015	0.154	0.015	0.940	0.000	0.000	1.508	0.62		63
	0.100	0.100	0.000	0.000	0.000	0.000	0.100	0.296	1.397	0.477	0.000	0.000	0.000	0.502	0.000	2.974	0.17		30
	0.033	0.000	0.016	0.000	0.000	0.000	0.000	0.000	0.107	0.016	0.000	0.000	0.000	0.261	0.911	1.344	0.68		57
	8.909	4.862	1.971	3.906	0.273	0.374	0.821	25.590	13.974	11.112	8.234	14.587	2.000	1.673	1.715	100.001			
	0.94	0.56	0.61	0.37	0.20	0.69	0.03	0.84	0.34	0.68	0.49	0.53	0.47	0.30	0.53				
	0.02	0.06	0.09	0.06	0.10	0.21	0.01	0.02	0.03	0.04	0.05	0.04	0.09	0.10	0.10				
	92	70	50	95	34	29	28	280	156	126	93	159	54	38	55				1359

Level I

10	20	30	40	80	90	Total	Users	S.E.	*n*
3.768	0.057	0.114	0.057	0.057	0.066			0.02	84
0.009	4.681	0.131	0.444	1.219	0.035			0.03	183
0.009	0.037	1.987	1.376	1.072	0.009			0.12	125
0.265	1.873	5.530	40.339	7.596	1.747			0.02	592
0.081	1.689	1.110	2.795	17.762	0.389			0.03	292
0.144	0.093	0.159	1.859	0.168	1.277			0.07	83
8.814	10.674	1.282	51.216	24.831	3.183				
0.94	0.69	0.27	0.90	0.74	0.53				
0.02	0.04	0.10	0.01	0.03	0.08				
91	216	89	566	307	90				1359

Table 5
Error matrix for Level I and Level II classification: Mid-Atlantic region

Level II

11	21	22	23	31	32	33	41	42	43	81	82	85	91	92	Total	Users	S.E.	*n*
3.768	0.000	0.000	0.057	0.000	0.000	0.114	0.000	0.000	0.057	0.000	0.057	0.000	0.000	0.066	4.119	0.92		73
0.000	1.865	0.173	0.679	0.000	0.000	0.104	0.173	0.000	0.139	0.262	0.069	0.666	0.035	0.000	4.165	0.45		76
0.000	0.352	0.122	0.390	0.000	0.000	0.017	0.006	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.906	0.14		70
0.009	0.027	0.009	0.662	0.000	0.000	0.009	0.126	0.000	0.000	0.000	0.000	0.150	0.000	0.000	0.993	0.67		65
0.000	0.048	0.000	0.000	0.002	0.000	0.002	0.000	0.000	0.388	0.000	0.000	0.000	0.000	0.000	0.440	0.01		42
0.000	0.007	0.000	0.007	0.000	0.899	0.482	0.137	0.000	0.402	0.526	0.022	0.007	0.000	0.085	2.575	0.35		77
0.009	0.018	0.000	0.000	0.000	0.000	0.734	0.062	0.334	0.053	0.245	0.272	0.000	0.000	0.009	1.735	0.42		66
0.000	1.523	0.000	0.000	0.000	0.761	4.328	18.944	1.217	7.615	2.512	1.655	0.767	0.761	0.000	40.084	0.47		118
0.000	0.132	0.000	0.085	0.000	0.000	0.557	0.978	2.476	1.472	0.340	0.085	0.000	0.255	0.000	6.380	0.39		66
0.265	0.000	0.132	0.132	0.000	0.000	0.132	2.302	2.139	3.196	0.000	0.586	0.397	0.645	0.000	9.927	0.32		68
0.000	0.455	0.000	0.683	0.000	0.000	0.455	1.223	0.228	0.455	4.538	7.504	1.291	0.228	0.000	17.061	0.27		77
0.081	0.161	0.000	0.161	0.000	0.000	0.242	0.322	0.308	0.242	0.242	3.919	0.461	0.081	0.081	6.301	0.62		64
0.000	0.005	0.394	0.061	0.000	0.000	0.033	0.005	0.005	0.005	0.000	0.423	0.682	0.000	0.000	1.615	0.42		79
0.000	0.050	0.000	0.000	0.000	0.000	0.151	1.264	0.126	0.418	0.050	0.075	0.000	0.680	0.000	2.814	0.24		79
0.144	0.034	0.000	0.000	0.000	0.000	0.008	0.017	0.017	0.017	0.008	0.017	0.025	0.025	0.572	0.886	0.65		84
4.265	4.679	0.831	2.918	0.002	1.660	7.368	25.560	6.849	14.460	8.725	14.685	4.465	2.710	0.810	100.001			
0.88	0.40	0.15	0.23	1.00	0.54	0.10	0.74	0.36	0.22	0.52	0.27	0.15	0.25	0.71				
0.05	0.11	0.09	0.06	0.00	0.23	0.05	0.04	0.07	0.10	0.10	0.04	0.08	0.08	0.13				
79	88	33	101	6	23	98	158	68	112	54	107	76	40	61				1104

Level I

Class	10	20	30	40	80	90	Total	Users	
3.768	0.057	0.114	0.057	0.057	0.066			0.04	73
0.009	4.681	0.131	0.444	1.219	0.035			0.06	217
0.009	0.037	1.987	1.376	1.072	0.009			0.12	181
0.265	1.873	5.530	40.339	7.596	1.747			0.04	257
0.081	1.689	1.110	2.795	17.762	0.389			0.04	213
0.144	0.093	0.159	1.859	0.168	1.277			0.07	163
4.275	8.428	9.031	46.869	27.875	3.522				
0.88	0.56	0.22	0.86	0.64	0.36				
0.05	0.08	0.08	0.02	0.05	0.09				
79	222	127	338	237	101				1104

Table 6
Error matrix for Level I and Level II classification: Southeast region

Level II

11	21	22	23	31	32	33	41	42	43	81	82	85	91	92	Total	Users	S.E.	*n*
4.718	0.000	0.000	0.000	0.000	0.000	0.047	0.000	0.047	0.000	0.000	0.000	0.012	0.000	0.047	4.871	0.97		101
0.000	1.639	0.025	0.051	0.044	0.000	0.000	0.019	0.019	0.172	0.000	0.000	0.248	0.019	0.019	2.255	0.73		115
0.000	0.164	0.157	0.092	0.012	0.000	0.000	0.000	0.000	0.006	0.006	0.000	0.030	0.000	0.000	0.467	0.34		72
0.023	0.233	0.012	0.569	0.151	0.000	0.012	0.012	0.012	0.310	0.023	0.023	0.095	0.012	0.012	1.499	0.38		89
0.100	0.112	0.000	0.100	4.125	0.100	0.000	0.000	0.000	0.000	1.204	0.602	1.505	0.000	0.702	8.550	0.48		87
0.012	0.000	0.000	0.021	0.074	0.180	0.037	0.029	0.000	0.021	0.017	0.000	0.012	0.008	0.004	0.415	0.43		97
0.000	0.037	0.000	0.000	0.000	0.000	1.344	0.153	0.102	0.330	0.179	0.077	0.327	0.236	0.051	2.836	0.47		85
0.047	0.056	0.000	0.000	0.000	0.000	1.046	8.789	0.000	2.670	0.251	0.127	0.422	0.248	0.124	13.780	0.64		132
0.000	0.025	0.000	0.238	0.238	0.000	0.026	0.238	11.680	6.536	0.124	0.000	0.238	1.585	0.133	21.061	0.56		107
0.000	0.218	0.000	0.000	0.000	0.000	0.697	0.533	0.026	13.738	0.325	0.000	0.348	0.174	0.000	16.059	0.86		111
0.000	0.022	0.000	0.000	0.000	0.000	0.001	0.243	0.001	0.029	0.702	0.431	0.054	0.000	0.000	1.483	0.47		95
0.047	0.001	0.000	0.000	0.102	0.000	0.003	0.006	0.000	0.154	0.241	0.686	0.342	0.029	0.112	1.723	0.40		111
0.000	0.197	0.102	0.051	0.000	0.000	0.000	0.000	0.000	0.045	0.178	0.076	1.163	0.025	0.000	1.837	0.63		74
0.180	0.000	0.000	0.000	0.108	0.000	0.432	0.540	0.349	1.138	0.208	0.216	0.108	7.799	0.432	11.510	0.68		110
0.133	0.000	0.133	0.133	0.277	0.000	0.000	0.000	0.000	0.796	0.265	0.000	0.133	0.796	8.992	11.658	0.77		88
5.260	2.704	0.429	1.255	5.131	0.280	3.645	10.562	12.236	25.945	3.723	2.238	5.037	10.931	10.628	100.004			
0.90	0.61	0.37	0.45	0.80	0.64	0.37	0.83	0.96	0.53	0.19	0.31	0.23	0.71	0.85				
0.04	0.07	0.15	0.12	0.08	0.24	0.08	0.05	0.02	0.04	0.07	0.11	0.04	0.05	0.05				
109	147	34	67	78	41	72	113	68	204	119	105	123	106	88				1474

Level I

10	20	30	40	80	90	Total	Users	S.E.	
4.088	0.000	0.047	0.047	0.000	0.047	4.228	0.97	0.02	93
0.144	3.202	0.188	0.521	0.344	0.073	4.472	0.72	0.06	274
0.109	0.239	5.984	0.595	3.810	0.793	11.529	0.52	0.07	264
0.105	0.556	1.158	47.024	1.354	1.628	51.825	0.91	0.02	364
0.093	0.377	0.103	0.106	4.073	0.129	4.881	0.83	0.04	279
0.313	0.133	0.817	1.788	1.063	18.952	23.065	0.82	0.03	200
4.852	4.505	8.296	50.081	10.643	21.622	100.000			
0.84	0.71	0.72	0.94	0.38	0.88				
0.05	0.07	0.07	0.01	0.05	0.03				
104	248	188	392	347	195				1474

Table 7
Regional comparison of class-specific accuracy, mode definition of agreement

Class	User's accuracy (%)				Producer's accuracy (%)			
		NY/NJ	Mid-Atlantic	Southeast		NY/NJ	Mid-Atlantic	Southeast
(a) Level II								
11	97	98	92	97	92	94	88	90
21	48	72	45	73	50	56	40	61
22	39	57	14	34	60	61	15	37
23	43	72	67	38	40	37	23	45
31	29	15	1	48	2	20	100	80
32	35	53	35	43	12	69	54	64
33	52	7	42	47	24	3	10	37
41	48	61	47	64	61	84	74	83
42	38	92	39	56	35	34	36	96
43	39	55	32	86	49	68	22	53
81	48	42	27	47	17	49	52	19
82	32	62	62	40	70	53	27	31
85	24	62	42	63	19	47	15	23
91	27	17	24	68	44	30	25	71
92	32	68	65	77	22	53	71	85
(b) Level I								
10	97	98	91	97	92	94	88	84
20	69	87	72	72	66	69	56	71
30	56	35	44	52	13	27	22	72
40	82	84	70	91	95	90	86	94
80	59	79	75	83	53	74	64	38
90	55	42	35	82	58.	53	36	88

inclusion probability of pixel u, incorporating both stages of sampling (Särndal et al., 1992, Chapter 9), was

$$\pi_{hu} = \pi_{2.1hu}\pi_{1u} = (n_h/N_h^*)(1/K). \quad (1)$$

In some cases, pixels selected for the sample were not interpreted (e.g., no photo could be located). These pixels were regarded as missing at random, and n_h in the formula for π_{hu} was revised to the number of sample pixels in the stratum for which reference data were obtained.

The NY/NJ region was analyzed in the same way, but the underlying rationale for the analysis requires different justification. In this region, a supplemental sample was employed to augment the general sample to increase the sample size for several rare classes. The supplemental rare-class sample was chosen from the same first-stage sample PSUs selected by the general NY/NJ design. For the general sample, a post-stratified analysis (Särndal et al., 1992, Sections 7.6 and 7.10.2) was employed in which the poststratified, conditional inclusion probabilities were $\pi_{2.1hu} = n_h'/N_h^*$, where n_h' is the number of sample pixels of class h observed in the general design. For the supplemented rare classes, the data from the two samples were combined. The sample size (n_h) used to compute π_{hu} (Eq. (1)) for the rare-class stratum h was the total number of pixels in the sample from both the general and rare-class designs. This derivation of π_{hu} is justified by recognizing that the poststratified, conditional inclusion probability from the general design is that of a stratified random sample, and the pixels brought in from the rare-class design are also selected by a stratified random protocol. A stratified random sample augmented by another stratified

Table 8
Estimated accuracy (%) using center pixel and mode agreement definitions, and homogeneous map land-cover subset: New England region

Class	User's accuracy			Producer's accuracy		
	Center	Mode	Homogeneous	Center	Mode	Homogeneous
(a) Level II (n = 1573 for center and mode definitions, n = 413 for homogeneous subset)						
11	99	97	100	92	92	77
21	48	48	60	53	50	97
22	35	39	39	55	60	11
23	46	43	61	38	40	89
31	28	29	69	3	2	67
32	27	35	60	12	12	100
33	46	52	40	24	24	62
41	48	48	58	66	61	81
42	41	38	59	34	35	41
43	37	39	33	46	49	25
51	58	75	85	4	3	19
81	50	48	91	16	17	40
82	24	32	19	48	70	46
85	24	24	30	24	19	50
91	29	27	24	56	44	52
92	27	32	76	24	22	87
Overall	46	46	56	–	–	–
(b) Level I (n = 1573 for center and mode definitions, n = 676 for homogeneous subset)						
10	99	97	100	92	92	77
20	63	69	79	66	66	89
30	50	56	62	17	13	50
40	82	82	94	93	95	97
50	58	74	85	4	3	12
80	56	59	88	48	53	80
90	41	55	71	69	58	97
Overall	78	80	90	–	–	–

random sample, both selected from the same first-stage sample PSUs, is still stratified random, and the appropriate inclusion probability is $\pi_{2.1hu} = n_h/N_h^*$ (Rao, 1985).

The inclusion probabilities are known for all pixels in the sample, and they are greater than zero for all pixels in the mapping region. These two conditions establish the probability sampling basis of the design. Eq. (1) also shows that within each mapping region, all pixels mapped as Level II land-cover class h had the same inclusion probability. Stratified random sampling formulas were applied to estimate the error matrix and associated summary measures. We next develop these general estimation formulas.

Let $y_{hu}(i,j)$ be the observation recorded for sample pixel u, where the h subscript indicates that pixel u was selected from stratum h. Define $y_{hu}(i,j) = 1$ if the agreement definition results in pixel u belonging to map class i and reference class j in the error matrix; otherwise, $y_{hu}(i,j) = 0$ (i.e., pixel u does not fall into cell (i,j) of the error matrix). Note that i and j may refer either to an Anderson Level I or Level II class, but h is always a Level II class determined by the original stratification. The value of $y_{hu}(i,j)$ depends on the definition of agreement employed. The estimation weight associated with pixel u is the reciprocal of the inclusion probability,

$$w_{hu} = \frac{1}{\pi_{hu}} = \frac{KN_h^*}{n_h} \tag{2}$$

w_{hu} is not affected by the definition of agreement because it is determined by the sampling design, not the response design.

Estimating the parameter N_{ij}, the number of pixels in the region that belong to cell (i,j) of the error matrix, serves as the starting point of the analysis:

$$\hat{N}_{ij} = \sum_S w_{hu} y_{hu}(i,j), \tag{3}$$

where $\sum_S$ indicates summation over all sample pixels. Once $\hat{N}_{ij}$ is calculated for all cells of the error matrix, we can construct the following estimators (the parameter estimated is listed in parentheses and q denotes the number of land-cover classes):

$$\hat{N} = \sum_i^q \sum_j^q \hat{N}_{ij} \quad \text{(total number of pixels mapped)} \tag{4}$$

$$\hat{p}_{ij} = \hat{N}_{ij}/\hat{N}$$

(proportion of pixels in cell$(i,j)$of the error matrix) (5)

$$\hat{p}_{i+} = \sum_j^q \hat{p}_{ij} \quad \text{(proportion of pixels mapped as class } i\text{)}$$

$$\hat{p}_{+j} = \sum_i^q \hat{p}_{ij} \tag{6}$$

(proportion of pixels belonging to reference class j) (7)

$$\hat{P} = \sum_i^q \hat{p}_{ii} \quad \text{(overall proportion correctly classified)} \tag{8}$$

Table 9
Estimated accuracy (%) using center pixel and mode agreement definitions, and homogeneous map land-cover subset: New York/New Jersey region

Class	User's accuracy				Producer's accuracy			
	CenterP	Center	Mode	Homogeneous	CenterP	Center	Mode	Homogeneous
(a) Level II (n = 1359 for center and mode definitions, n = 463 for homogeneous subset)								
11	94	94	98	100	91	93	94	100
21	25	59	72	83	17	56	56	83
22	50	52	57	62	18	29	61	88
23	63	67	72	100	16	20	37	62
31	77	79	15	75	12	20	20	52
32	64	64	53	53	23	18	69	63
33	40	40	7	54	3	3	3	88
41	49	57	61	71	67	77	84	98
42	80	84	92	94	25	33	34	22
43	18	44	55	71	38	68	68	
81	30	33	42	46	61	65	49	
82	59	64	62	86	23	26	53	
85	61	63	62	100	24	27	47	93
91	14	18	17	11	27	30	30	10
92	67	69	68	71	28	35	53	91
Overall	47	55	62	76	–	–	–	
(b) Level I (n = 1359 for center and mode, n = 838 for homogeneous subset)								
10	94	94	98	100	91	94	94	94
20	72	83	87	95	45	55	69	68
30	63	63	35	61	9	12	27	94
40	76	80	84	89	85	89	90	97
80	69	76	79	89	66	71	74	60
90	31	34	42	74	34	41	53	52
Overall	74	79	82	88	–	–	–	

The estimates in column CenterP are based on agreement defined as a match between the map label and primary reference label.

$\hat{p}_{ii}/\hat{p}_{i+}$ (user' s accuracy for map class i) (9)

$\hat{p}_{jj}/\hat{p}_{+j}$ (producer' s accuracy for reference class j) (10)

These general formulas possess the flexibility to accommodate different definitions of agreement (i.e., simply change $y_{hu}(i,j)$ to correspond to the definition desired), and they are applicable for estimating accuracy parameters for various subsets of the sample. For the simplest case in which the stratum class h corresponds to the map class i, the general formulas reduce to familiar stratified sampling formulas (Card, 1982). However, the formulas also apply to those definitions of agreement for which the stratum class, h, and row of the error matrix, i, differ. For example, for the agreement definition based on the mode, suppose the map label of the sample pixel (i.e., center pixel in the support region) is class 81 (pasture/hay), but the mode class of the support region is class 82 (row crops). The estimation weight (w_{hu}) of that pixel is determined by π_{hu} for the class 81 stratum, but the map class of this sample pixel is class 85 in the error matrix cell entry. The inclusion probability is identified by the class 81 map label because it is this center pixel label, not the mode label, that is used in the sample selection protocol. Changing a pixel's map label after the sample has been selected does not retroactively change π_{hu} because π_{hu} is determined at the time the sample is selected.

5.2. Variance estimation

Särndal et al. (1992, Section 9.4) discuss a general approach to variance estimation applicable to two-stage sampling designs. For the NLCD analyses, a variance approximation (see Appendix A) is employed permitting use of estimation procedures available in Statistical Analysis Software (SAS, Version 8.2, 2001). The approximation treats the design as a cluster sample in which each PSU contains pixels of only a single map class. These PSUs are then sampled by a stratified random design. The SAS variance approximation treats pixels from different strata as belonging to different clusters, even though in the actual NLCD design, pixels from different strata may fall within the same PSU. The SAS variance algorithm estimates only the among-PSU (first-stage) variance component of the two-stage design. This approximation, typically employed in sampling practice, is justified by the dominance of the among-PSU variance component over the smaller within-PSU (or second-stage) variance component (Särndal et al., 1992, Section 4.3). Because the accuracy estimators (Eqs. (8)–(10)) are ratio estimators, a Taylor series linearization (Särndal et al., 1992, Section 5.5) is used in the variance approximation. The variance approximation incorporates several key design features, specifically the stratification by mapped land-cover class, the estimation weights, and the clustering structure. The SAS variance algorithm does not treat sample pixels within each cluster as independent, thereby improving the variance estimator over using simple random or stratified random variance estimation formulas.

Because practical exigencies made it difficult to unambiguously recover the PSU identities of all sample pixels in each of these four mapping regions, we used map polygons instead of the design PSUs as clusters in the variance approximation. A map polygon is a collection of contiguous pixels, all of the same land cover. Map polygons are defined based on the Level II classification. To ensure that replacing the PSUs with the map polygons had no significant effect on the estimated standard errors, we compared the standard errors computed from both approaches using data collected for three mapping regions outside the eastern U.S. for which we had access to both PSU and map polygon information. In all three of these other mapping regions, negligible differences arose between the standard errors derived from clusters defined as the PSU and those derived from the map polygon clusters. Incorporating the cluster structure in the standard errors resulted in higher standard errors, as expected, than standard errors derived from a binomial model, which assumes independent sample observations. The comparison of variance estimators is shown for one region in the table of Appendix A.

6. Results

For the mode definition of agreement, estimated overall accuracy for the Level II classification ranges from a low of

Table 10
Estimated accuracy (%) using center pixel and mode agreement definitions, and homogeneous map land-cover subset: Mid-Atlantic region

Class	User's accuracy			Producer's accuracy		
	Center	Mode	Homogeneous	Center	Mode	Homogeneous
(a) Level II (n = 1104 for center and mode definitions, n = 377 for homogeneous subset)						
11	87	92	100	88	88	100
21	36	45	46	21	40	77
22	27	14	13	13	15	24
23	67	67	94	13	23	60
31	15	1	14	100	100	100
32	27	35	39	8	54	100
33	41	42	42	3	10	10
41	40	47	69	76	74	79
42	33	39	70	27	36	92
43	21	32	40	15	22	15
81	27	27	32	57	52	71
82	58	62	67	20	27	20
85	49	42	61	2	15	11
91	31	24	43	21	25	46
92	62	65	88	57	71	80
Overall	38	43	61	–	–	–
(b) Level I (n = 1104 for center and mode, n = 579 for homogeneous subset)						
10	87	91	100	88	88	96
20	67	72	76	29	56	78
30	44	44	49	5	22	10
40	65	70	84	88	86	93
80	73	75	87	63	64	85
90	41	35	55	31	36	40
Overall	67	70	84	–	–	–

46% in the Mid-Atlantic region to a high of 66% in the Southeast region. For the Level I classification, overall accuracy ranges from 70% in the Mid-Atlantic to 83% in the Southeast (Table 2). Standard errors for estimated overall accuracy range between 1% and 4%.

The estimated error matrices, user's and producer's accuracies, and standard errors are provided in Tables 3–6 to document the regional, class-specific accuracy for Levels I and II. Approximate 95% confidence intervals for the summary accuracy parameters may be constructed using the estimated accuracy ± 2 standard errors. The usual interpretation of an error matrix applies: the diagonal elements display information on correct classifications, and the off-diagonal elements identify misclassifications. Rather than attempt to summarize the detailed information encapsulated in the error matrices, we leave it to the reader to extract accuracy information relevant to an intended application of the 1992 NLCD. User's and producer's accuracies are organized by region in Table 7 to facilitate regional comparisons of class-specific accuracies.

Because accuracy depends on the definition of agreement, class-specific accuracy estimates are provided for several other settings in Tables 8–11. In the first setting, the map label of only the sample pixel is used, as opposed to the 3×3 map support region used in the mode definition of agreement (see results in Tables 3–7). Accuracy estimates constructed on this basis do not take into account confounding attributable to misregistration between the reference and map locations. Agreement is defined as a match between the map class and either the primary or alternate reference label (column labeled Center in Tables 8–11). For the NY/NJ and Southeast regions, results are also reported for a second definition of agreement (column labeled CenterP in Tables 9 and 11), where a match is declared if the map label matches the primary reference label (the alternate reference label, if present, is ignored). This is the strictest definition of agreement, permitting no allowance for location uncertainty or thematic ambiguity. In the New England and Mid-Atlantic regions, few pixels were assigned alternate labels, so results based on using both the primary and alternate reference labels differ little from those using only the primary label. The final set of results (column labeled Homogeneous in Tables 8–11) are derived from the subset of sample pixels for which all pixels in the 3×3 map support region consist of a single map class (i.e., homogeneous map land cover). Accuracy estimates derived from this homogeneous subset are largely immune from the effects of spatial misregistration. Results for the mode definition of agreement used to construct the full error matrices are included for comparison.

Accuracy results based on the center pixel definition of agreement (match with primary or alternate reference label) are slightly lower than those estimated from the mode agreement definition (Tables 8–11). For the regional overall

Table 11
Estimated accuracy (%) using center pixel and mode agreement definitions, and homogeneous map land-cover subset: Southeast region

Class	User's accuracy				Producer's accuracy			
		Center	Mode	Homogeneous	CenterP	Center	Mode	Homogeneous
(a) Level II (n = 1474 for center and mode definitions, n = 469 for homogeneous subset)								
11	89	94	97	100	76	88	90	96
21	46	58	73	38	30	39	61	55
22	11	28	34	23	14	30	37	29
23	33	40	38	65	32	39	45	54
31	33	45	48	57	74	84	80	96
32	35	38	43	48	32	58	64	100
33	29	36	47	40	17	31	37	50
41	48	65	64	86	64	76	83	93
42	34	46	56	80	79	88	96	97
43	64	80	86	100	38	51	53	24
81	29	55	47	69	1	2	19	3
82	57	63	40	82	3	4	31	11
85	41	59	63	50	18	28	23	23
91	43	68	68	77	46	64	71	94
92	61	69	77	96	77	83	85	94
Overall	48	62	66	81	–	–	–	–
(b) Level I (n = 1474 for center and mode, n = 713 for homogeneous subset)								
10	89	94	97	100	76	88	84	96
20	57	69	72	83	46	59	71	87
30	36	46	52	50	49	74	72	69
40	79	90	91	94	86	92	94	96
80	58	72	83	95	13	19	38	16
90	64	78	82	90	74	85	88	93
Overall	69	80	83	89	–	–	–	–

The estimates in column CenterP are based on agreement defined as a match between the map label and primary reference label.

estimates, the decrease in accuracy using this definition of agreement relative to the mode definition ranges from 0% to 7% at Level II and from 2% to 3% at Level I, and the class-specific accuracies are lower than the mode accuracies for most, but not all classes. The accuracies derived from the homogeneous subset are considerably higher than the accuracies based on the full sample. The increase in the class-specific accuracies relative to the mode definition ranges from 10% to 18% for Level II and from 6% to 14% for Level I. These findings are consistent with the general expectation that using only homogeneous areas of the map for accuracy assessment will overestimate accuracy (Hammond & Verbyla, 1996). For the NLCD, excluding heterogeneous areas omits a significant proportion of the mapped area. For the four regions, the percent of the sample represented by the homogeneous subset ranges from 26% to 34% at Level II and from 43% to 62% at Level I. Although not useful to characterize accuracy of the NLCD, the homogeneous subset results do provide quantitative insight into differences in error rates between edge versus interior pixels. Comparison of the CenterP and Center columns of Tables 9 and 11 illustrates the impact of thematic ambiguity on the accuracy estimates in these two regions.

7. Discussion

7.1. Sampling design

Multi-stage sampling offers flexibility to tailor an accuracy assessment sampling design for specific objectives. In the NLCD, the two-stage design achieved large-scale spatial control over the sample thus reducing aerial photography costs, and at the same time created a frame from which to select a sample stratified by mapped land-cover class. Consequently, two of the primary design criteria were satisfied: limiting the cost of reference data materials and ensuring adequate sample sizes for each map class to obtain reasonably precise estimates of user's accuracy. Nusser and Klaas (2003) used a similar two-stage structure, with USGS 7.5 quadrangles serving as the PSUs to achieve spatial control over the sample and reduce field costs. Their design also included provision for stratifying by land-cover class. Edwards, Moisen, and Cutler (1998) is another example in which a two-stage structure was effectively used to control the spatial distribution of the sample, while still allowing stratification, in this case by proximity to roads.

Because selection of the first-stage PSUs was not dependent on any characteristic of the map, the NLCD two-stage design may be adapted to other purposes. For example, these same first-stage sample PSUs could serve as the basis of a two-stage cluster sample for assessing the accuracy of a change product created from the 1992 NLCD and the proposed 2000 NLCD. Within the first-stage PSUs, pixels may be stratified by mapped change (Biging, Colby, & Congalton, 1998), and a stratified sample selected. An advantage of using the existing 1992 NLCD accuracy design is that the first-stage PSUs provide a legitimate frame from which a variety of design adaptations may be implemented depending on the specific objectives of an assessment. In the change detection application, a significant savings on aerial photography costs for the 1992 NLCD reference data could be achieved because much of the photography already purchased for the current assessment would cover sample locations selected in a change detection assessment based on the same design.

The class-specific standard errors show a general tendency for user's accuracy to be estimated more precisely than producer's accuracy. In some cases, the producer's accuracy standard error is high. This phenomenon is the result of signing the sample to satisfy the criterion of precise estimation of user's accuracies. User's accuracy precision is controlled by the sample allocation to the map land-cover class strata. The allocation chosen in the 1992 NLCD, however, creates highly variable estimation weights among strata. Estimating producer's accuracy requires combining data across strata, and the high variation of the stratum estimation weights may translate to large standard errors. Stratifying by mapped land-cover class represents a decision to favor precision of user's accuracy over precision of producer's accuracy. Although some producer's accuracy standard errors appear alarmingly high, this is a natural consequence of a design tailored to ensure precision of user's accuracy estimates.

7.2. Response design

Ensuring consistency of the response design protocol among different teams of interpreters is a challenging and costly activity in a large-area accuracy assessment. Consistent reference data collection protocols were implemented within each region. Although each region received the same general guidelines for the response design, there was flexibility to interpret the instructions in different ways. Because the response design protocol was not completely objective, some regional differences are expected. For example, the percent of the sample assigned an alternate reference label varied regionally (11.5% in New England, 33.6% in NY/NJ, and 56.9% in the Southeast), with no alternate reference labels provided in the Mid-Atlantic. Regional variability in the assignment of alternate labels would be expected even if the same interpreter team had assessed all four regions. However, it was apparent that the variation in use of alternate labels was partially responsible for the regional differences in estimated accuracy. Multiple teams of interpreters are a practical reality of such a large-scale assessment because of workload and timing considerations. Methods for achieving greater regional consistency in assignment of alternate labels should be developed. Alternatively, statistical calibration techniques will need to be developed to standardize estimates for differences attributable to variability in response design implementation.

The NLCD protocol defining agreement as a match between the map class and either the primary or alternate reference label produces results analogous to those of a fuzzy-class assessment. In a fuzzy approach (Gopal & Woodcock, 1994), a linguistic scale value is assigned to each land-cover class for each sample pixel, where the linguistic scale has 1 = absolutely wrong, 2 = understandable but wrong, 3 = reasonable or acceptable answer, 4 = good answer, and 5 = absolutely right. Agreement may be defined in several ways, most commonly via MAX and RIGHT operators. The MAX operator defines agreement to occur when the land-cover class with the highest linguistic scale reference value matches the map land-cover label. The RIGHT operator is a more liberal definition. An agreement occurs under the RIGHT operator when the land-cover class identified by the map has a linguistic scale value of, say, 3 or higher in the fuzzy reference data.

To relate the NLCD results to those of a fuzzy assessment, we first note that the primary and alternate reference labels used in the NLCD protocol can be viewed as representing the two highest linguistic scale values assigned via a fuzzy protocol. Although the exact linguistic values of these primary and alternate reference labels are unknown, the primary label represents the class having the maximum value on the linguistic scale, and the alternate label corresponds to the reference class with the second highest linguistic scale value. Consequently, the accuracy measures derived by defining agreement as a match between the map label and either the primary or alternate reference label are analogous to estimates obtained by the fuzzy class RIGHT operator. A definition of agreement using only the primary reference label would produce results akin to the fuzzy class MAX operator. The connection between the NLCD and fuzzy approach based on the RIGHT operator is further strengthened, if in the fuzzy approach, a high proportion of the sample has Membership 1 or 2, where Membership is defined for each pixel as the number of land-cover classes for which the linguistic scale value is 3 or above. For example, if only fuzzy Membership 1 or 2 values are present, then each sample pixel has, in effect, been assigned only a primary and one alternate label. Assessments in which Membership values are predominantly 1 or 2 have been reported by Muller et al. (1998) and Laba et al. (2002), with the former having 100% of the sample with Membership 1 or 2, and the latter having over 94% of the sample with Membership 1 or 2.

Although it is easier to communicate results for a single definition of agreement, in the absence of perfect reference data, it is unclear what this one definition should be. Reporting estimates for a variety of agreement definitions provides NLCD users the flexibility to choose a definition relevant to their application. Neither fuzzy set (Gopal & Woodcock, 1994) nor rough set (Ahlqvist, Keukelaar, & Oukbir, 2000, p. 483) approaches are constrained to a single definition of agreement. Examining multiple definitions of agreement also provides insight into the potential effects of various errors in the reference data, thereby further enhancing the information content of the accuracy assessment and our understanding of the map's utility.

Because the reference data are obtained from aerial photography, the estimated accuracies should be viewed as representing agreement with interpreted land cover. Reference data collection is expensive by practically any means implemented, and using aerial photography was the most practical strategy given the budget and magnitude of the NLCD assessment. Ground visits were considered, but ruled out as being cost-prohibitive. We also considered limited ground visits as a check on interpretation accuracy, but decided against this option because the number of ground visits required to provide an assessment beyond mere anecdotal evidence would still be prohibitive. Ground visits to only a few sample sites of each land-cover class would not likely provide better information than what was already being obtained by the quality control procedures in place for the photointerpretation procedures. The regional accuracy results suggest that some of the accuracy problems may be resolvable by implementing the interpretation protocol differently, and/or perhaps redoing some of the interpretations. However, given the budget available for accuracy assessment, such reanalysis is not probable. The reported accuracy estimates represent agreement as determined by the reference data collection protocol, and as such are the best estimates currently available.

7.3. Analysis

The accuracy estimators employed in the NLCD assessment are constructed to achieve the criterion of statistical consistency (Stehman, 2000). In practical terms, consistent estimation requires incorporating an appropriate weight for each sample pixel in the analysis. These weights are determined by the inclusion probabilities. Because of the stratified feature of the design, inclusion probabilities vary considerably by land-cover class. Consequently, the common practice of reporting error matrices in terms of sample counts cannot be followed here because these counts are not representative of the population. Instead, error matrices are reported as estimated cell proportions to properly characterize the map population.

A second important feature of the analysis is that standard errors are reported. Although it is conventional statistical practice to report a standard error for each estimate, this is not always done in accuracy assessment. Omitting standard errors may be attributable to the complexity of the variance estimation formulas, even for relatively simple designs such as stratified random sampling (without clustering). Lack of statistical software specifically tailored to accuracy assessment is another problem (Stehman & Czaplewski, 1998). The design-based, survey sampling estimation procedures in SAS

provided a convenient way to implement variance estimation for the complex design employed in the NLCD assessment.

The high cost of sampling for accuracy assessment can be ameliorated by more efficient, but more complex probability sampling designs such as that implemented in the NLCD, and the designs described in Nusser and Klaas (2003) and Edwards et al. (1998). However, more complex designs typically require more complex variance estimators. Estimating precision is a lower priority criterion relative to the cost and class-specific accuracy criteria specified for the NLCD design. Achieving cost-effectiveness by accepting an approximate, rather than unbiased estimator of variance is a reasonable compromise given that many accuracy assessment studies do not report standard errors at all.

8. Summary

The objectives of this article were to document the accuracy assessment methodology and report results for the four mapping regions comprising the eastern United States. Regional error matrices are provided to document accuracy for both Levels I and II of the 1992 NLCD, and accuracy estimates are provided for two different agreement definitions as well as for a subset of the sample consisting of homogeneous 3×3 blocks of mapped land cover. The sampling design met the specified criteria of being a probability sample, ensuring precision of the user's accuracy estimates via stratification by land-cover class, and achieving cost-effectiveness via the two-stage cluster structure. Reporting standard errors for the estimated accuracies is another key development. Eight of the 10 1992 NLCD mapping regions now have accuracy assessments completed, and assessments of the other 2 regions are in progress. Upon completion of these two regions, the accuracy results for the national 1992 NLCD map will be reported.

Acknowledgements

We thank an anonymous reviewer for very helpful and detailed constructive comments. The high quality reference data used in these analyses were provided by the North Carolina State Center for Earth Observation, via EPA Contract 9V1031NAEE, Versar, via EPA Contract 68-W60023, and Lockheed Martin Corporation, via EPA Contract 68-C50065. Ray Czaplewski and Zhiliang Zhu were instrumental in the design of the NY/NJ assessment. This article has not been subject to EPA or USGS review and does not necessarily reflect the views of either agency. Mention of trade names is not intended as an endorsement for any particular product or vendor.

Appendix A

The variance estimation formula used in the SAS SURVEYMEANS procedure is described below. The notation used in this appendix follows that used in the SAS program documentation, and therefore should be viewed independently of notation used elsewhere in the text. The following notation is required: w_{hij} is the reciprocal of the inclusion probability (weight) for pixel j in cluster i of stratum h; y_{hij} is an indicator variable for pixel j in cluster i of stratum h (e.g., y_{hij} is 1 if the pixel belongs to row r and column c of the error matrix, 0 otherwise); x_{hij} is also an indicator variable for pixel j in cluster i of stratum h (e.g., x_{hij} is 1 if the pixel belongs to column c of the error matrix, 0 otherwise).

The ratio estimator

$$\hat{R} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}x_{hij}}$$

can be used to estimate the different accuracy metrics by proper choice of x_{hij} and y_{hij}. For example, if $y_{hij}=1$ when pixel j's map and reference labels place it on the diagonal entry of the error matrix for class K ($y_{hij}=0$ otherwise) and $x_{hij}=1$ if pixel j has reference class K ($x_{hij}=0$ otherwise), $\hat{R}$ estimates producer's accuracy of class K. The estimated variance of $\hat{R}$ is then

$$\mathrm{var}(\hat{R}) = \sum_{h=1}^{H} \frac{n_h(1-n_h/N_h)}{(n_h-1)} \sum_{i=1}^{n_h} (g_{hi.} - \bar{g}_{h..})^2,$$

where

$$g_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - x_{hij}\hat{R})}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}x_{hij}} \quad \text{and} \quad \bar{g}_{h..} = \left(\sum_{i=1}^{n_h} g_{hi.}\right) \Big/ n_h$$

Table A: Comparison of standard errors based on using map polygons and the original PSUs as the clusters in the variance approximation. SEPatch denotes a standard error using the map polygon (patch) as the cluster and SEPSU denotes a standard error using the original PSU as the cluster. The column SEBino has results using the binomial standard error formula, $\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p}$ is the estimated user's accuracy and n is the number of sample pixels for that class. Because producer's accuracy combines data over strata, the simple binomial standard error formula is not available. The last two columns show the difference in standard errors between using a patch versus a PSU as the

CL	Est. users	S.E. for user's accuracy				SE producer's		Difference in S.E. SEPatch-SEPSU	
		SEPatch	SEPSU	SEBino	n	SEPatch	SEPSU	Users	Prod.
11	0.96				33				
21	0.85				91				
22	0.24				99				
23	0.49				94				
31	0.59				89				
32	0.33				98				
33	0.56				92				
41	0.61				96				
42	0.63				94				
43	0.75				92				
51	0.52				93				
81	0.33				94				
82	0.16				94				
83	0.90				98				
84	0.64				97				
85	0.57				95				
91	0.43				92				
92	0.35				93				
All	0.64				1634				

cluster. The data are from EPA Federal Region 5 in the midwestern U.S.

References

Ahlqvist, O., Keukelaar, J., & Oukbir, K. (2000). Rough classification and accuracy assessment. *International Journal of Geographical Information Science, 14*, 475–496.

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). A land use and land cover classification system for use with remote sensor data. *U.S. Geological Survey Prof. Paper 964*. Washington, DC: U.S. Geological Survey, 28 pp.

Biging, G. S., Colby, D. R., & Congalton, R. G. (1998). Sampling systems for change detection accuracy assessment. In R. S. Lunetta, & C. D. Elvidge (Eds.), *Remote sensing change detection: Environmental monitoring methods and applications* (pp. 281–308). Chelsea, MI: Ann Arbor Press.

Card, D. H. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing, 48*, 431–439.

Edwards Jr., T. C., Moisen, G. G., & Cutler, D. R. (1998). Assessing map accuracy in a remotely-sensed ecoregion-scale cover-map. *Remote Sensing of Environment, 63*, 73–83.

Gopal, S., & Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing, 60*, 181–188.

Hammond, T. O., & Verbyla, D. L. (1996). Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing, 17*, 1261–1266.

Laba, M., Gregory, S. K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J., & DeGloria, S. D. (2002). Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover maps. *Remote Sensing of Environment, 81*, 443–455.

Lanter, D. P., & Veregin, H. (1992). A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing, 58*, 825–833.

Loveland, T. R., & Shaw, D. M. (1996). Multiresolution land characterization: Building collaborative partnerships. In T. Tear, & M. Scott (Eds.), *Gap analysis, a landscape approach to biodiversity planning* (pp. 17–25). Bethesda, MD: American Society of Photogrammetry and Remote Sensing.

Muller, S. V., Walker, D. A., Nelson, F. E., Auerbach, N. A., Bockheim, J. G., Guyer, S., & Sherba, D. (1998). Accuracy assessment of a land-cover map of the Kuparuk River Basin, Alaska: Considerations for remote regions. *Photogrammetric Engineering and Remote Sensing, 64*, 619–628.

Nusser, S. M., & Klaas, E. E. (2003). Survey methods for assessing land cover map accuracy. *Environmental and Ecological Statistics*, (to appear).

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology, 11*, 15–31.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model-assisted survey sampling*. New York: Springer-Verlag.

Stehman, S. V. (2000). Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment, 72*, 35–45.

Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering and Remote Sensing, 67*, 727–734.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment, 64*, 331–344.

Stehman, S. V., Wickham, J. D., Yang, L., & Smith, J. H. (2000). Assessing the accuracy of large-area land cover maps: Experiences from the Multi-resolution Land-Cover Characteristics (MRLC) project. In G. B. M. Heuvelink, & M. J. P. M. Lemmens (Eds.), *Accuracy 2000: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 601–608). The Netherlands: Delft Univ. Press.

Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., & Van Driel, N. (2001). Completion of the 1990s national land cover data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing, 67*, 650–662.

Yang, L., Stehman, S. V., Smith, J. H., & Wickham, J. D. (2001). Thematic accuracy of MRLC land cover for the eastern United States. *Remote Sensing of Environment, 76*, 418–422.

Yang, L., Stehman, S. V., Wickham, J. D., Smith, J. H., & VanDriel, N. J. (2000). Thematic validation of land cover data of the eastern United States using aerial photography: Feasibility and challenges. In G. B. M.

zaplewski, & M. J. P. M. Lemmens (Eds.), *Accuracy 2000: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 747–754). The Netherlands: Delft Univ. Press.

Zhu, Z., Yang, L., Stehman, S. V., & Czaplewski, R. L. (1999). Designing an accuracy assessment for a USGS regional land cover mapping program. In K. Lowell, & A. Jaton (Eds.), *Spatial accuracy assessment: Land information uncertainty in natural resources* (pp. 393–398). Chelsea, MI: Ann Arbor Press.

Zhu, Z., Yang, L., Stehman, S. V., & Czaplewski, R. L. (2000). Accuracy assessment for the U.S. Geological Survey regional land-cover mapping program: New York and New Jersey region. *Photogrammetric Engineering and Remote Sensing, 66*, 1425–1435.